Alzheimer's & Dementia: Translational Research & Clinical Interventions 3 (2017) 219-228

Featured Article

Predicting mild cognitive impairment from spontaneous spoken utterances

Meysam Asgari[a,*], Jeffrey Kaye[b], Hiroko Dodge[b,c]

[a]Center for Spoken Language Understanding, Oregon Health & Science University (OHSU), Portland, Oregon, USA
[b]Department of Neurology, Layton Aging and Alzheimer's Disease Center, Oregon Health & Science University (OHSU), Portland, Oregon, USA
[c]Department of Neurology, Michigan Alzheimer's Disease Center, University of Michigan, Ann Arbor, Michigan, USA

Abstract

Introduction: Trials in Alzheimer's disease are increasingly focusing on prevention in asymptomatic individuals. We hypothesized that indicators of mild cognitive impairment (MCI) may be present in the content of spoken language in older adults and be useful in distinguishing those with MCI from those who are cognitively intact. To test this hypothesis, we performed linguistic analyses of spoken words in participants with MCI and those with intact cognition participating in a clinical trial.

Methods: Data came from a randomized controlled behavioral clinical trial to examine the effect of unstructured conversation on cognitive function among older adults with either normal cognition or MCI (ClinicalTrials.gov: NCT01571427). Unstructured conversations (but with standardized preselected topics across subjects) were recorded between interviewers and interviewees during the intervention sessions of the trial from 14 MCI and 27 cognitively intact participants. From the transcription of interviewees recordings, we grouped spoken words using Linguistic Inquiry and Word Count (LIWC), a structured table of words, which categorizes 2500 words into 68 different word subcategories such as positive and negative words, fillers, and physical states. The number of words in each LIWC word subcategory constructed a vector of 68 dimensions representing the linguistic features of each subject. We used support vector machine and random forest classifiers to distinguish MCI from cognitively intact participants.

Results: MCI participants were distinguished from those with intact cognition using linguistic features obtained by LIWC with 84% classification accuracy which is well above chance 60%.

Discussion: Linguistic analyses of spoken language may be a powerful tool in distinguishing MCI subjects from those with intact cognition. Further studies to assess whether spoken language derived measures could detect changes in cognitive functions in clinical trials are warrented.

Keywords: Biomarkers; Conversational interactions; Early identification; Mild cognitive impairment (MCI); Social markers; Speech characteristics

1. Introduction and motivation

A well-documented literature has identified characteristic early disruption of normative patterns and processing of speech and language in patients with Alzheimer's disease (AD) as well as in prodromal dementia states such as mild cognitive impairment (MCI) [1]. Early foundational clinical studies of language have highlighted changes in verbal fluency and naming [2–4]. More recent studies using automated or semiautomated speech and language analysis approaches have identified linguistic as well as acoustic features that characterize early AD or MCI such as pause frequency and duration, and linguistic complexity measures [5,6].

Almost all of these latter studies have used elicited speech paradigms to generate speech and language samples, for

*Corresponding author. Tel.: 503-346-3752; Fax: 503-346-3754.
E-mail address: asgari@ohsu.edu

http://dx.doi.org/10.1016/j.trci.2017.01.006

example, asking patients to describe what they observe in pictures briefly presented to them or to recall specific stories they are exposed to during a testing session. In addition to analyzing the conversations in these structured, mostly constrained within a clinical setting, there are some studies which have used more spontaneous speech [7,8]. In spite of the potential advantages of capturing spontaneous speech in conversations, major barriers have existed to implementing this approach for persons with MCI or AD in more natural settings. A major impediment has been limitations in the recording technology paradigms that could be deployed. This has been both a problem of practicality such as the form factor of recording devices and power requirements for long-term recording, as well as automated speech and linguistic analysis challenges. Despite these challenges, pioneering early studies using somewhat obtrusive worn or carried recording devices have shown the potential power of this approach in younger populations. For example, Pennebaker and Mehl have illustrated the value of inferring social contexts from audio life logs using a lexicon of salient words, termed Linguistic Inquiry and Word Count (LIWC) 2001 [9]. They demonstrated that social context and other information from audio life logs can be used to quantify participants' social life (interaction and engagement), cognitive function, emotional conditions, and even health status [10]. To the best of our knowledge, LIWC analyses have not been used to examine the cognitive status of older adults. In this study, we use LIWC on a corpus of spontaneous speech samples generated during the course of a 6-week randomized clinical trial of daily online video chats to improve social engagement and cognition in older adults with and without MCI [11,12]. These conversations between the interviewer and the participant provided an opportunity to analyze potential differences in the conversational output of persons with MCI and cognitively intact adults.

1.1. Language and mild cognitive impairment

Although the most typical early cognitive deficit observed in Alzheimer's disease involves the memory domain, linguistic ability is also clearly affected. For example, secondary verbs per utterance, percentage of clauses, percentage of right-branching and left-branching clauses, propositions per utterance, conjunctions per utterance, mean duration of pauses, and standardized phonation time have all been reported to show significant differences between healthy older adults and subjects with MCI or AD [5,13–20]. A major barrier to taking advantage of these language-based discriminators has been the effort required to manually score relevant features from speech samples; the proposed work addresses this through automatic scoring.

1.1.1. Related computational works

Recently, there has been considerable interest in automatically analyzing acoustic and language properties of speech samples to create more sensitive quantitative assessments of patients with cognitive impairment [1,21–23]. For example, Jerrold and colleagues [24] evaluated the ability of machine learning methods to differentiate dementia subtypes, including AD, based on semistructured conversational speech recordings. Their proposed method uses both acoustic features such as duration of consonants, vowels, and pauses, as well as lexical features such as frequency of nouns and verbs derived from automatic transcriptions provided by a speaker-independent automatic speech recognition (ASR) system.

Combining these two profiles of features derived from 48 participants, including nine healthy controls, nine AD patients, and 30 frontotemporal lobar degeneration (FTLD) patients (nine with behavioral variant frontotemporal dementia, 13 with semantic dementia, and eight with progressive nonfluent aphasia), they obtained 61% accuracy in detecting the subjects' FTLD subtype, significantly better than the random diagnosis condition, which had 20% accuracy. In a binary classification setting, they obtained 88% accuracy in distinguishing nine participants with AD from nine healthy controls. Similarly, Lehr et al. [25] developed an automated assessment system and applied it to spoken responses of subjects on a delayed recall test (Wechsler Logical Memory test). First, they automatically transcribed the recordings using an ASR system, then they extracted the story elements using the Berkeley aligner [26], and finally they compared those to the story elements manually identified by the expert examiner. Using a support vector machine (SVM) classifier applied to 72 participants, they showed ASR-derived features can distinguish 35 participants with MCI from 37 healthy controls with a classification accuracy of 81%. More recently, Toth et al. [27] presented an automatic approach for detecting MCI from speech samples in which participants were asked to talk about a 1-minute long animated film. They used an ASR system to transcribe the recordings and extract acoustic biomarkers including articulation rate, speech tempo, length of utterance, duration, and number of silent and filled pauses (hesitation). Their results show that the SVM classifier trained on the aforementioned acoustic features can distinguish 32 participants with MCI from 19 healthy controls with an accuracy of about 80%. Based on this prior work, we sought to improve the ability to extract meaningful markers of cognitive change from the spontaneous speech of individuals with MCI or those at risk for MCI.

2. Methods

2.1. Data collection and corpus

The present study was a part of a larger randomized controlled clinical trial that assessed whether frequent conversations conducted via webcam and Internet-enabled personal computers could improve cognitive function in older persons with either normal cognition

or MCI (ClinicalTirals.gov registration number: NCT01571427). The study protocol and the results have been described in detail elsewhere [12]. Briefly, in the larger intervention trial, social interaction sessions were conducted using semistructured conversations with trained interviewers for 30 minutes a day, 5 days a week for 6 weeks (i.e., 30 sessions) among the intervention group; the control group did not have daily video-chat sessions. Inclusion and exclusion criteria are listed in Table 1. There was high adherence to the daily video-chat protocol (89%; range, 77%–100%).

The Clinical Dementia Rating (CDR) scale [28] was used to classify participants into groups and defined MCI as CDR 0.5 and cognitively intact as CDR 0. Out of 41 participants randomized to the intervention group, 33 consented to allow their daily conversational intervention sessions to be transcribed for speech characteristic analyses (21 cognitively intact; 12 MCI). In addition, eight participants (six cognitively intact, two MCI) recruited during a pilot-testing study who went through the same intervention protocol also consented and were included in the study described here, resulting in a total of 41 participants. The transcribed interviewees' speech during their daily chat sessions over 6 weeks was analyzed in this study. Table 2 reports the baseline characteristics of participants including age, education, gender, and Mini–Mental State Examination scores.

Table 1
Inclusion and exclusion criteria used in the trial

Inclusion criteria
1. Age 70 years or older
2. CDR = 0 or 0.5
3. Sufficient vision and hearing to engage in conversation by personal computer system.
4. Sufficient English language skills to complete all testing.
5. General health status that will not interfere with ability to complete longitudinal study. Conditions that will likely lead to this problem are listed in the following in the study exclusions list.

Exclusion criteria
1. Plan to start taking new classes, traveling which requires more than two nights of stay away, or having significant social events such as a family wedding or a family reunion, during the scheduled prevention trial.
2. Diseases associated with dementia such as AD, ischemic vascular dementia, normal pressure hydrocephalus, or Parkinson's disease.
3. Significant disease of the central nervous system such as brain tumor, seizure disorder, subdural hematoma, cranial arteritis.
4. Current (within the last 2 years) alcohol or substance abuse
5. Current major depression, schizophrenia, or other major psychiatric disorder
6. Unstable or significantly symptomatic cardiovascular disease such as coronary artery disease with frequent angina, or congestive heart failure with shortness of breath at rest.
7. Active systemic cancer within 5 years of study entry.
8. Illness that requires >1 visit per month to a clinician.
9. Progressive vision loss (age-related macular degeneration already beginning to significantly degrade vision).
10. Need for oxygen supplementation for adequate function.
11. Medications:
    A. Frequent use of high doses of analgesics.
    B. Sedative medications except for those used occasionally for sleep (use limited to no more than twice per week).
    C. Applicable to CDR = 0.5 group only: subjects on unstable dosing of cholinesterase inhibitors (need to be stable dosing for 2 months).

Abbreviations: CDR, Clinical Dementia Rating; AD, Alzheimer's disease.

Table 2
Baseline characteristics of participants

Variable	Intact, $n = 27$	MCI, $n = 14$	P-value
Age	78.9 (5.5)	83.4 (8.8)	.10
Gender (% women)	63	86	.17
Years of education	16.6 (2.4)	14.0 (2.6)	.003
MMSE	28.7 (1.3)	26.9 (2.1)	.008

Abbreviations: MCI, mild cognitive impairment; MMSE, Mini–Mental State Examination.

2.2. Language analysis using LIWC

Our proposed method explores automating the identification of individuals with MCI using computational analysis of narrative language samples. We extract linguistic features using LIWC from manual transcription of unstructured conversations between interviewers and participating older adults as indicators of how participants and interviewers interact during the conversation.

LIWC2001, which we refer to as LIWC in this study, includes more than 2500 words or word stems categorized into groups of words known as "word subcategories" that tap a particular cluster of related words (e.g., negative emotion words). There are 68 word subcategories in LIWC each titled with a representative term that generates an overall "subcategory scale." For example, a group of job-related words such as "Employ," "Boss," and "Career" are grouped into a word subcategory of "Occupation." These word subcategories further cluster into five broader domains termed "word categories": (1) Linguistic Dimensions, (2) Psychological Processes, (3) Relativity, (4) Personal Concerns, and (5) Spoken Categories [9]. Each of these broad word categories includes words that represent a particular conceptual domain; for example, Linguistic Dimensions groups all personal and impersonal pronouns. Psychological processes denotes affective or emotional categories of words such as "Positive Emotion" and "Negative Emotion" subcategories, as well as cognitive processes such as a "Causation" subcategory, and social processes such as "Family" and "Friends" subcategories. Relativity includes a group of words that denote "Time" such as the tense of verbs, "Space," and "Motion." Personal Concerns includes a group of categories associated with personal matters such as occupation, financial issues, and so forth. Finally, the Spoken Categories class includes three categories of "Swear Words" such as *crap* and *goddam*, "Nonfluencies" such as *hm* and *umm*, and "Fillers" such as *youknow*. Table 3 provides a comprehensive list of the default LIWC word categories, subcategory scales,

Table 3
LIWC2001 output variable information

Category	Subcategory scale	Examples	Count of words
Linguistic processes	Total pronouns	I, our, they	70
	1st person singular	I, me, my	9
	1st person plural	we, us, our	11
	Total 1st person	I, we, me	20
	Total 2nd person	you, you'll	14
	Total 3rd person	she, their, them	22
	Negations	no, not, never	31
	Assent	agree, OK, yes	18
	Articles	a, an, the	3
	Prepositions	to, with, above	43
	Numbers	Second, thousand	34
Personal concerns	Occupation	work, class, boss	213
	School	class, student, college	100
	Job	employ, boss, career	62
	Achievement	goal, hero, win	60
	Leisure activity	TV, chat, movie	102
	Home	apartment, kitchen	26
	Sports	football, game, play	28
	TV and movies	TV, sitcom, cinema	19
	Music	tunes, song, CD	31
	Money	income, cash, owe	75
	Metaphysical	God, church, coffin	85
	Religion	altar, church, mosque	56
	Death	dead, coffin, kill	29
	Physical states	ache, breast, sleep	285
	Body states	ache, heart, cough	200
	Sex and sexuality	lust, penis, fuck	49
	Eating	eat, swallow, taste	52
	Sleeping	sleep, bed, dreams	21
	Grooming	wash, bath, clean	15
Psychological processes	Affective	happy, ugly, bitter	6
	Positive emotion	happy, pretty, good	261
	Positive feelings	happy, joy, love	43
	Optimism	Certainty, pride, win	69
	Negative emotion	hate, worthless, enemy	345
	Anxiety	nervous, afraid, tense	62
	Anger	hate, kill, pissed	121
	Sadness	grief, cry, sad	72
	Cognitive process	cause, know, ought	312
	Causation	because, effect, hence	49
	Insight	think, know, consider	116
	Discrepancy	should, would, could	32
	Inhibition	block, constrain	64
	Tentative	maybe, perhaps, guess	79
	Certainty	always, never	30
	Sensory process	see, touch, listen	111
	Seeing	view, saw, look	31
	Hearing	heard, listen, sound	36
	Feeling	touch, hold, felt	30
	Social process	talk, us, friend	314
	Communication	talk, share, converse	124
	Other references	1st, 2nd, 3rd	54
	Friends	pal, buddy, coworker	28
	Family	mom, brother, cousin	43
	Humans	boy, woman, group	43
Relativity	Time	hour, day, o'clock	113
	Past verb	walked, were, had	144
	Present verb	walk, is, be	256
	Future verb	will, might, shall	14
	Space	around, over, up	71
	Up	up, above, over	12
	Down	down, below, under	7
	Inclusive	with, and, include	16
	Exclusive	but, except, without	19
	Motion	walk, move, go	73
Spoken categories	Swear words	damn, fuck, piss	29
	Nonfluencies	uh, rr*	6
	Fillers	youknow, Imean	6

Abbreviation: LIWC, Linguistic Inquiry and Word Count.

NOTE. List of the default LIWC word categories (first column), subcategory scales (second column), a few examples from word subcategories (third column), and frequency of words found in each word subcategory (forth column).

examples from word subcategories, and count of words that exist in each word subcategory. The selection of words defining the LIWC categories involved multiple steps over several years, initially, to collect groups of words representing basic emotional and cognitive dimensions. Here, we briefly review the development steps of LIWC and refer readers to the LIWC user's manual [9] for more detail.

First, sets of words were generated for each word subcategory. Next, using several sources, such as the positive and negative affect scales [29] for the Psychological Processes word category, relevant words were generated by a group of 3–6 judges for all word subcategories. Then, three independent judges indicated whether each suggested word properly fits within its word subcategory. Words for which judges could not decide on appropriate category placement were discarded. A majority voting among judges determined final candidates to each word subcategory. Percentages of agreement for judges ratings were acceptable for all LIWC word subcategories ranging from a low of 86% agreement for the subcategory of "Optimism" to 100% agreement for the subcategory of "Humans." One should note that each word or word stem may be part of one or more word subcategories in LIWC. For example, the word "cried" is part of four word subcategories: "Sadness," "Negative Emotion," "Affective," and "Past Tense Verb." Detailed information about LIWC word categories can be found in [30]. LIWC has been widely used in a range of applications, and its reliability has been validated for a range of problems such as linguistic analysis of social media [31] or analyzing and discovering personality traits [32]. In this study, we use LIWC to automatically distinguish participants with MCI from healthy controls using linguistic features extracted from the content of spontaneous conversation.

Our linguistic analysis of transcriptions began with grouping spoken words into 68 LIWC word subcategories. Note that raw transcriptions are stemmed before splitting into word categories. The stemming process refers to

extracting the stem or root of words so that words with the same roots such as "book" and "books" fall into the same word subcategory. Next, we count the number of words that fall into each word subcategory. This generates a vector of 68 dimensions referring to 68 word counts on each word subcategory. Some words may not belong to any of LIWC word categories and these are discarded; 39.8% of words were found unclassifiable to any of the 68 word categories. Because the total number of words spoken by participants at interview sessions is not equal, the dynamic range of features may vary among participants and this may confound classification performance. To address this issue, we normalize each count by dividing it by the total number of words. We treat this vector as an input feature vector to our classification algorithm. Moreover, to study the relative importance of each group of the five word categories for distinguishing participants with MCI from those with intact cognition, we train five different classifiers each with linguistic features derived only from one of the main groups of word categories in a secondary analysis. Fig. 1 represents the block diagram of our proposed method for extracting and modeling linguistic features of participants' transcriptions to distinguish participants with MCI from those with intact cognition. In the next section, we present the learning strategies and experimental setup.

2.3. Learning strategies

To explore the effectiveness of different learning methods in distinguishing participants with MCI from those with intact cognition, we trained statistical models based on extracted linguistic features using two widely employed machine learning algorithms: (1) SVM [33] and (2) random forest classifier (RFC) [34].

2.3.1. Problem definition

Distinguishing participants with MCI from those with intact cognition can be cast into a hypothesis test problem, in which true and null hypotheses, H_1 and H_0, are the prediction of the participant as MCI and cognitively intact, respectively. Given a set of linguistic features derived from the transcription of conversational interviews, one must first train statistical models of MCI and cognitively intact classes that well represent inherent characteristics of both H_1 and H_0 hypotheses. The efficiency of this process, known as model training, depends on the quality of extracted features as well as the discriminant power of the learning algorithm that separates classes with the highest margin. Let the D-dimensional linguistic feature vectors (D is total number of features) extracted from the transcription of a participant be x_i and $y_i \in \{+1,-1\}$ his or her class label where 1 and -1 represent the participant's cognitive status (MCI vs. intact). Thus, we need to learn a classification function $f(x)$, that predicts the subject's label from the available training data $D = (x_i,y_i)$; $i = 1,\ldots,n$, where n is the total number of participants.

2.3.2. Classification algorithms

SVMs [33] are among the best supervised learning methods widely used in pattern recognition, classification, and regression problems. A SVM classifier constructs a hyperplane in a high-dimensional space to best discriminate data points belonging to different classes. In nonlinear cases,

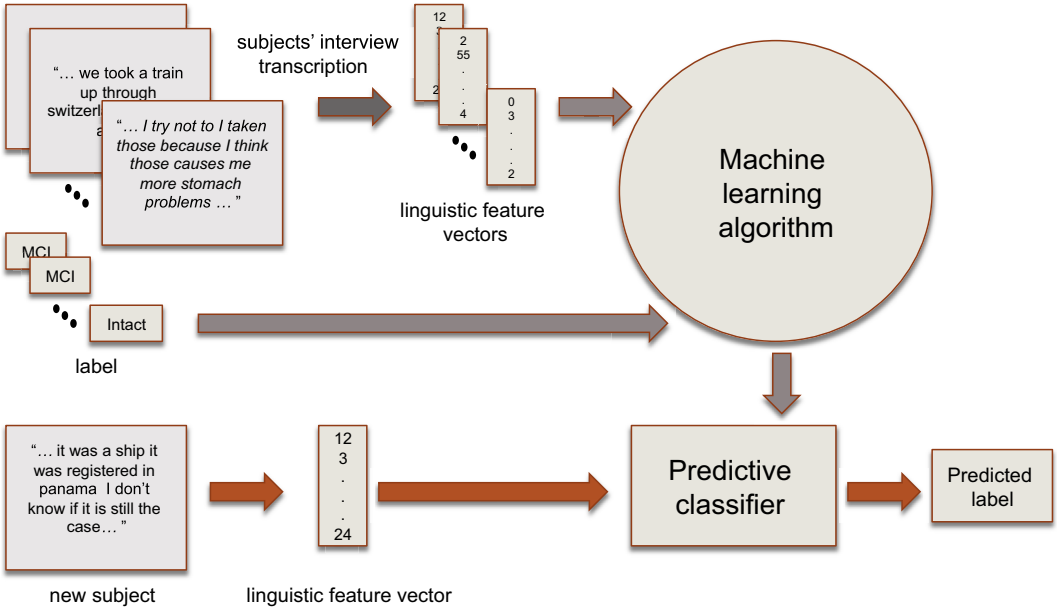

Fig. 1. Block diagram of extracting and modeling linguistic features of participants' transcriptions to distinguish participants with MCI from those with intact cognition. Abbreviation: MCI, mild cognitive impairment.

SVM leverages from a mathematical technique called Kernel trick [35] to first map input features into a high-dimensional space and then find a hyperplane that maximizes the class margin. One of the main advantages of SVM is its effectiveness in cases where the dimension of feature vectors is greater than the number of training samples. This makes the use of SVM particularly suitable in our experiment where there is a relatively small pool of subjects versus the minimum feature dimension of 68. We train a linear SVM classifier as well as a nonlinear SVM classifier with a radial basis function (RBF) kernel used from the open-source Scikit-learn toolkit [36] independently for different sets of features from LIWC. In the machine learning literature [35], a nonlinear SVM is recommended for classifying data points that are not linearly separable. We also use an RFC [34] that trains a number of decision tree classifiers on randomly drawn subsamples of the data set and then combine these decision tree classifiers to improve the predictive accuracy and to control overfitting. When a new input sample is entered into the RFC, it is first classified by all of the decision trees and then voting majority criteria will estimate the class label of the input sample. The simplest classification function, referred to as "Chance" in our experiments, is a random classifier that corresponds to randomly classifying all subjects into two classes.

2.3.3. LIWC features

Before describing our different modeling strategies, we outline the feature extraction procedure in this task. Transcriptions of the recordings were produced by nonprofessional transcribers via the Amazon Mechanical Turk (AMT) crowd-sourcing platform. To assess the quality of transcriptions, we randomly picked four interview sessions (125 minutes total) and evaluated the word error rate (WER) between transcriptions provided by AMT nonprofessional and professional transcribers. The WER that measured the percentage of deleted, inserted, and substituted words in the AMT transcriptions with respect to the reference (gold standard) professional transcriptions was approximately 15% suggesting an acceptable agreement between AMT-derived and reference transcription. From the transcription of interviewees recordings, we grouped spoken words using the LIWC lexicon into 68 different subcategories and counted the number of words grouped in each subcategory of LIWC as a representative feature. This resulted in a 68-dimensional feature vector representing the linguistic information of each participant. As noted earlier, each feature was normalized by the total word count of each participant. In addition, we examined the relative importance of each main group of word categories of LIWC in our classification problem.

2.3.4. Cross-validation

To validate how our statistical analyses and experimental results were independent of our data sets, we used cross-validation (CV) techniques in which the train and test sets are rotated over the entire data set. We used a five-fold cross-validation scheme, setting all model parameters using four of the sets as the training set, and using the fifth one only for reporting the performance estimates. Parameters of the optimal SVM model were determined on the training set separately for each fold via grid search and cross-validation.

2.3.5. Performance criteria

To evaluate the performance of the proposed classifier, we adopted the following evaluation metrics: (1) Accuracy—in our binary classification, accuracy is the proportion of participants that are correctly identified in both intact and MCI classes divided by the total number of participants. The accuracy itself does not represent the performance of the model due to the imbalanced number of participants in our cohort; (2) Sensitivity—the portion of correctly identified MCI participants (true positives). Sensitivity (SE) assesses the capability of the model to distinguish MCI from cognitively intact participants; (3) Specificity—the portion of correctly identified cognitively intact participants (true negative). Specificity (SP) measures how well the model is at avoiding false positives; (4) Area under the curve of receiver operating characteristics (AUC-ROC)—the most common method for evaluating the performance of a binary classifier is the receiver operating characteristics [37], which plots the sensitivity (true positive rate) of the classifier versus 1 − specificity (false positive rate) of the classifier as the classification threshold varies. We use a classification threshold in a grid search schema to cover the most positive threshold (everything true) to the most negative threshold (everything false). In our experimental setup, we report the average over five iterations of the CV for every performance criteria.

2.3.6. Imbalanced data

Because of the imbalanced number of participants in our cohort, partitioning data into train and test sets via CV could result in an imbalanced test set. For example, in a five-fold scenario, randomly assigning 20% of 41 participants, 14 with MCI and 27 cognitively intact, into the test set might result in a case where one MCI participant coincides with seven cognitively intact participants in the test set. This will result in a highly imbalanced test set in which performing CV will negatively affect the overall conclusion on the performance of the classifier. We tackle this potential issue through an iterative process. First, we randomly permute the entire data, perform five-fold CV, and accumulate averaged scores at the end of each iteration. Next, we calculate the overall performance by taking the average of 5-fold CV scores across iterations. The iteration is repeated until the overall performance converges to a steady state. Our experiments showed that after about 200 iterations of 5-fold CV, the overall performance converged.

Table 4
Comparison of performance of different classifiers distinguishing participants with MCI from those with intact cognition

Classifier	Sensitivity	Specificity	Accuracy	AUC-ROC
Chance	30.0	76.0	60.0	52.2
Nonlinear SVM (RBF)	53.2	88.2	76.2	71.2†
Linear SVM	60.96	77.5	71.9	69.2†
Linear SVM + L1-norm	72.7	72.4	72.4	72.5†
RFC	6.51	72.3	74.7	68.2†

Abbreviations: MCI, mild cognitive impairment; AUC-ROC, area under the curve of receiver operating characteristics; SVM, support vector machine; RBF, radial basis function; RFC, random forest classifier.
NOTE. †$P < .05$.

3. Results

Using features extracted from the LIWC lexicon, we compared the performance of the aforementioned classifiers for distinguishing participants with MCI from those cognitively intact controls. As discussed earlier, the number of linguistic features extracted from transcriptions here is larger than the number of participants. Given this scenario, the learning task is an ill-posed problem without a unique solution for the linear function. The simplest solution to this problem is to automatically eliminate those features that are not informative. This can be performed by augmenting the cost function of the SVM classifier with a regularization term that penalizes large values of the regression coefficients, driving them to zero when they are not useful. In our experiments, we used a L1-norm regularization term that is well known in applications requiring sparse solutions, assigning zero values to useless regression coefficients [38].

The results are reported in Table 4 for the five-fold cross-validations. The performance of SVM classifiers with linear and RBF kernels as well as RFC in terms of Sensitivity, Specificity, Accuracy, and AUC-ROC are shown. We also repeated the experiment using a Chance classifier which randomly assigned participants into MCI and intact classes. Results indicate that all SVM classifiers and RFC perform significantly better than "Chance" (*P*-value of <.05), according to the cross-validated paired t-test [39] († denotes statistically significant results). As it is shown in Table 4, the SVM model with linear kernel and a L1-norm regularization term outperforms nonlinear SVM with an RBF kernel as well as the RFC in terms of AUC-ROC.

3.1. Effectiveness of LIWC dimensions in extracted linguistic features

Word categories in the LIWC2001 are generally arranged hierarchically, composed of five main classes of word categories: Personal Concerns, Relativity, Psychological Processes, Linguistic Dimensions, and Spoken Categories. In our initial experiment, we simply grouped spoken words into 68 LIWC word categories, and the resulting 68-dimensional linguistic features were used for learning MCI and intact SVM models. To study the relative importance of each group of the five word categories for distinguishing participants with MCI from intact volunteers, we used five different SVM models each with linguistic features derived only from one of the main groups of word categories in a secondary analysis.

The results for five-fold cross-validation are reported in Table 5 for the SVM model with the linear kernel and L1-norm regularization term. In this analysis, linguistic features extracted from the Linguistic Dimensions category by itself are not particularly effective at this task. Features from Spoken Categories are also not informative and underperform noticeably compared to other classes of features. This might be due to the small size of the feature set (only three features) derived from this category. In contrast, results show that features derived from Psychological Processes and Personal Concerns significantly outperformed the "Chance" classifier. Features from the Relativity class are best at distinguishing participants with MCI with sensitivity of 81% and AUC-ROC of 80%. Interestingly, this

Table 5
Comparison of performance using linguistic features extracted from five LIWC main groups of word categories, for distinguishing MCI subjects

LIWC categories	Number of features	Sensitivity	Specificity	Accuracy	AUC-ROC
Linguistic dimensions	17	64.37	55.43	69.0	62.2
Chance	17	30.7	76.3	60.9	53.5
Psychological processes	25	63.93	67.8	62.12	64.96†
Chance	25	32.1	76.9	61.8	54.5
Relativity	10	80.77	75.83	83.33	**79.61**†
Chance	10	30.6	76.2	60.9	53.4
Personal concerns	19	70.3	62.60	74.60	68.30†
Chance	19	30.1	76.1	60.7	53.1
Spoken categories	3	43.45	67.23	59.11	55.34
Chance	3	30.7	76.3	60.9	53.5

Abbreviations: LIWC, Linguistic Inquiry and Word Count; MCI, mild cognitive impairment; AUC-ROC, area under the curve of receiver operating characteristics (best result indicated in bold).
NOTE. †$P < .05$.

Table 6
Baseline characteristics of subsampled participants

Variable	Intact, $n = 15$	MCI, $n = 14$	P-value
Age	79.4 (5.1)	83.4 (8.8)	.15
Gender (% women)	63%	86%	.17
Years of education	14.8 (1.37)	14.0 (2.6)	.31

Abbreviation: MCI, mild cognitive impairment.

category alone noticeably outperforms the system in which all 68 features are used.

3.2. Influence of education level

According to Table 2, there was a significant difference in the years of education between participants with MCI and those who are cognitively intact. It is possible that the level of education may significantly influence verbal abilities regardless of cognitive decline [40]. To control for education, we repeated the analysis with a subset of participants from the intact group that better matches the education level of participants in the MCI group. Table 6 reports the baseline characteristics of subsampled participants of more equal educational level and results are shown in Table 7. In this secondary analysis, the classifier trained on the features from the Relativity word category outperformed other classifiers. In addition, comparing results with those previously shown in Table 5 suggests that education plays a significant role in this study. Finally, we performed an analysis using a Student t-test on the averaged percentage of words that fall into the Relativity word category across spoken words of all participants from both MCI and intact classes. MCI participants used significantly more words ($P < .001$) than intact participants from word subcategories of the Relativity word category. This also indicates that MCI participants use more "verbs" than healthy controls according to Table 3. One of our speculations in this regard is that complex sentences could involve more words that just verbs (articles, adjectives, etc.) and therefore, more number of verbs indicate that sentence complexity is simpler among the MCI subjects. However, because of limited literature to support our hypothesis, we can not provide any in-depth explanation.

4. Discussion

In summary, we have reported our experiments on distinguishing MCI from cognitively intact older adults solely from the spontaneous speech recorded from conversational engagement sessions held for 41 study participants. Our results show that MCI participants can be distinguished from cognitively intact older adults with an accuracy of 84% using LIWC-driven features. Interestingly, combining all features from 68 word categories resulted in poorer performance suggesting that some word categories in the LIWC are not suitable for this task. We found that the linguistic features derived from word subcategories belonging to the Relativity word category are significantly better at capturing cues with MCI participants as compared to other classes in the LIWC lexicon and give the best classification results. However, this study is not able to explain the cognitive basis for the high performance achieved by the Relativity word category achieve high performance in this task. The linguistic approach used here could be applied to preclinical trials where enriching the study cohort with high-risk subjects and more sensitive outcomes to change are required. Standardized linguistic analysis of spontaneous conversations has the advantage of providing a measure of cognitive function that is inherently person-specific, conveniently captured and ecologically more valid than commonly used constrained psychometric testing sessions However, despite this promise, a current important limitation to this approach is that the analysis relies on high-fidelity transcription of the conversations which is labor intensive. However, we anticipate that there will be continued major advances in the accuracy of automated speech recognition, and thus, this methodology could be widely adopted in clinical practice to screen or identify those at risk of MCI and/or dementia in communities as well as for monitoring progression of disease. In addition to improvements in automated speech recognition for data capture, considerable work remains to improve

Table 7
Distinguishing 14 MCI from 15 cognitively intact participants with characteristics reported in Table 6 using LIWC feature sets

LIWC categories	Number of features	Sensitivity	Specificity	Accuracy	AUC-ROC
Chance	68	57.31	46.46	51.47	51.89
All categories	68	77.55	47.23	61.81	62.39†
Linguistic dimensions	17	59.85	55.13	65.28	57.49
Psychological processes	25	69.93	37.0	52.73	53.46
Relativity	10	74.23	78.70	76.42	**76.46**†
Personal concerns	19	65.21	51.83	58.11	58.52
Spoken categories	3	47.8	51.7	44.45	49.75

Abbreviations: MCI, mild cognitive impairment; LIWC, Linguistic Inquiry and Word Count; AUC-ROC, area under the curve of receiver operating characteristics (best result indicated in bold).
NOTE. $^{\dagger}P < .05$.

accuracy of the classification algorithms. Our linguistic analysis did not incorporate many other potentially useful features, relying entirely on the LIWC feature set. This approach ignores sentence structure and other contextual information. The word-based approach using LIWC misses the word context and would miss cases such as "not too bad." A valuable avenue for future research would be to explore the feasibility of natural language processing techniques to address this drawback using more sophisticated methods of linguistic analysis. Furthermore, when applying this approach in clinical trials or to the general population, one would typically add other potentially predictive features to the classification model such as age, gender, education, and family history of dementia. Future studies will need to examine larger and more diverse populations over time and explore the possible cognitive bases behind the findings of the present study.

Acknowledgments

This work was supported by NIH National Institute on Aging awards R01 AG033581, R01 AG042191, P30 AG008017, P30 AG024978.

RESEARCH IN CONTEXT

1. Systematic review: Early cognitive deficit observed in Alzheimer's affects linguistic ability. Indicators of mild cognitive impairment (MCI) may be present in the content of spoken language in older adults and can be useful in distinguishing those with MCI from those who are cognitively intact.
2. Interpretation: We performed linguistic analysis of spoken words to classify 14 participants with mild cognitive impairment (MCI) from 26 with intact cognition. Applying support vector machine classifier on extracted linguistic features, we classified MCI participants with accuracy of 84%, well above the chance, 60%.
3. Future direction: The linguistic approach used here could be applied to preclinical trials where enriching the study cohort with high-risk subjects and more sensitive outcomes to change are required.

References

[1] Taler V, Phillips NA. Language performance in Alzheimer's disease and mild cognitive impairment: A comparative review. J Clin Exp Neuropsychol 2008;30:501–56.
[2] Tombaugh TN. Trail making test a and b: normative data stratified by age and education. Arch Clin Neuropsychol 2004;19:203–14.
[3] Murphy KJ, Rich JB, Troyer AK. Verbal fluency patterns in amnestic mild cognitive impairment are characteristic of Alzheimer's type dementia. J Int Neuropsychol Soc 2006;12:570–4.
[4] Wechsler D. Wechsler adult intelligence scale–fourth edition (wais–iv). London: Pearson; 2014.
[5] Fraser KC, Meltzer JA, Rudzicz F. Linguistic features identify Alzheimers disease in narrative speech. J Alzheimers Dis 2015; 49:407–22.
[6] Roark B, Mitchell M, Hollingshead K. Syntactic complexity measures for detecting mild cognitive impairment. In: Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing. Association for Computational Linguistics; 2007. p. 1–8.
[7] Forbes-McKay K, Shanks MF, Venneri A. Profiling spontaneous speech decline in Alzheimer's disease: A longitudinal study. Acta Neuropsychiatr 2013;25:320–7.
[8] Romero B, Kurz A. Deterioration of spontaneous speech in ad patients during a 1-year follow-up: Homogeneity of profiles and factors associated with progression. Dement Geriatr Cogn Disord 1996;7:35–40.
[9] Pennebaker JW, Francis ME, Booth RJ. Linguistic Inquiry and Word Count: Liwc 2001. Mahwah: Lawrence Erlbaum Associates; 2001; 71: p. 2001.
[10] Mehl MR, Gosling SD, Pennebaker JW. Personality in its natural habitat: manifestations and implicit folk theories of personality in daily life. J Pers Soc Psychol 2006;90:862–77.
[11] Dodge HH, Mattek N, Gregor M, Bowman M, Seelye A, Ybarra O, et al. Social markers of mild cognitive impairment: Proportion of word counts in free conversational speech. Curr Alzheimer Res 2014;12:513–9.
[12] Dodge HH, Zhu J, Mattek NC, Bowman M, Ybarra O, Wild KV, et al. Web-enabled conversational interactions as a method to improve cognitive functions: Results of a 6-week randomized controlled trial. Alzheimers Dement (N Y) 2015;1:1–12.
[13] Kemper S, LaBarge E, Ferraro FR, Cheung H, Cheung H, Storandt M. On the preservation of syntax in Alzheimer's disease: Evidence from written sentences. Arch Neurol 1993;50:81–6.
[14] Lyons K, Kemper S, LaBarge E, Ferraro FR, Balota D, Storandt M. Oral language and Alzheimer's disease: A reduction in syntactic complexity. Aging Neuropsychol Cogn 1994;1:271–81.
[15] Bucks R, Singh S, Cuerden JM, Wilcock GK. Analysis of spontaneous, conversational speech in dementia of Alzheimer type: Evaluation of an objective technique for analysing lexical performance. Aphasiology 2000;14:71–91.
[16] Singh S, Bucks RS, Cuerden JM. Evaluation of an objective technique for analysing temporal variables in DAT spontaneous speech. Aphasiology 2001;15:571–83.
[17] Oulhaj A, Wilcock GK, Smith AD, de Jager CA. Predicting the time of conversion to mci in the elderly role of verbal expression and learning. Neurology 2009;73:1436–42.
[18] Fraser KC, Rudzicz F, Rochon E. Using text and acoustic features to diagnose progressive aphasia and its subtypes. In: INTERSPEECH. Citeseer; 2013. p. 2177–81.
[19] Clark D, Kapur P, Geldmacher D, Brockington J, Harrell L, DeRamus T, et al. Latent information in fluency lists predicts functional decline in persons at risk for Alzheimer disease. Cortex 2014; 55:202–18.
[20] Berisha V, Wang S, LaCross A, Liss J. Tracking discourse complexity preceding Alzheimer's disease diagnosis: A case study comparing the press conferences of Presidents Ronald Reagan and George Herbert Walker Bush. J Alzheimers Dis 2015;45:959–63.
[21] Ahmed S, Haigh AM, de Jager CA, Garrard P. Connected speech as a marker of disease progression in autopsy-proven Alzheimers disease. Brain 2013;136:3727–37.
[22] Kempler D. Language changes in dementia of the Alzheimer type. Dementia and Communication 1995;98–114.
[23] Salmon DP, Butters N, Chan AS. The deterioration of semantic memory in Alzheimer's disease. Can J Exp Psychol 1999;53:108–17.

[24] Jarrold W, Peintner B, Wilkins D, Vergryi D., Richey C, Gorno-Tempini ML, et al. Aided diagnosis of dementia type through computer-based analysis of spontaneous speech. In: Proceedings of the ACL Workshop on Computational Linguistics and Clinical Psychology. 2014. p. 27–36.

[25] Lehr M, Prud'hommeaux ET, Shafran I, Roark B. Fully automated neuropsychological assessment for detecting mild cognitive impairment. In: INTERSPEECH; 2012. p. 1039–42.

[26] Liang P, Taskar B, Klein D. Alignment by agreement. In: Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics: Association for Computational Linguistics. 2006. p. 104–111.

[27] Tóth L, Gosztolya G, Vincze V, Hoffmann I, Szatlóczki G. Automatic detection of mild cognitive impairment from spontaneous speech using asr. Dresden: ISCA; 2015.

[28] Morris JC, Ernesto C, Schafer K, Coats M, Leon S, Sano M, et al. Clinical dementia rating training and reliability in multicenter studies the Alzheimer's disease cooperative study experience. Neurology 1997; 48:1508–10.

[29] Watson D, Clark LA, Tellegen A. Development and validation of brief measures of positive and negative affect: The PANAS scales. J Pers Soc Psychol 1988;54:1063–70.

[30] Pennebaker JW, Booth RJ, Francis ME. Linguistic inquiry and word count: LIWC [computer software]. Austin, TX: liwc. net; 2007.

[31] Hutto CJ, Gilbert E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In: Eighth International AAAI Conference on Weblogs and Social Media. 2014.

[32] Schwartz HA, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones SM, Agrawal M, et al. Personality, gender, and age in the language of social media: The open-vocabulary approach. PLoS One 2013;8:e73791.

[33] Smola AJ, Schölkopf B. A tutorial on support vector regression. Stat Comput 2004;14:199–222.

[34] Breiman L. Random forests. Machine Learn 2001;45:5–32.

[35] Schölkopf B, Smola AJ. Learning With Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. Cambridge: MIT press; 2002.

[36] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in python. The J Machine Learn Res 2011;12:2825–30.

[37] Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982; 143:29–36.

[38] Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B (Methodol) 1996;58:267–88.

[39] Dietterich TG. Approximate statistical tests for comparing supervised classification learning algorithms. Neural Comput 1998;10:1895–923.

[40] Mathuranath P, George A, Cherian P, Alexander A, Sarma SG, Sarma PS. Effects of age, education and gender on verbal fluency. J Clin Exp Neuropsychol 2003;25:1057–64.